Date created 31/5/2006

Some Updates on Genetic Data Analysis

This session is rather short, seeing that I have just done presentations for the journal club and work in progress meeting. I will highlight three aspects as follow-up for earlier presentations at genetic meetings, briefly on the foreign language calls from SAS and Stata, more on retrospective models, and then other update.

1. Foreign language calls from SAS and Stata

I noted earlier (Zhao and Tan, 2006) to call C/C++/Fortran codes through system shells as well as SAS/Toolkit. In fact, there is possibility to replace codes from these foreign languages with Stata plugin available from Stata 8.1 (details available from http://www.stata.com/plugins/), which will improve proficiency on many occasions. It is known that SAS/IML implements a well-established and powerful matrix language, with many mathematical/statistical and graphical routines ready to use. There is now a separate Windows product called SAS/IML Workshop which implements the IMLPlus language to line up SAS procedures, SAS/IML and C/Fortran/Java, and available from SAS 8.2.

2. Retrospective methods

The equivalence of risk estimates from both retrospective and prospective methods in epidemiology was shown by Prentice (1976). A comparison of retrospective (e.g. Epstein and Satten 2003) and prospective likelihood (e.g. Schaid et al. 2002) methods was reported by Satten and Epstein (2004). Under multiplicative model, the two methods are roughly comparable. However, for dominant and recessive models of haplotype effect, the retrospective-likelihood method has increased efficiency with respect to the prospective methods. The prospective methods are robust to departure from Hardy-Weinberg equilibrium, while the retrospective-likelihood method is biased for dominant and recessive models of haplotype effect. A more recent synthesis was given by Lin and Zeng (2006).

For haplotype analysis, Tan et al. (2005) pointed out that prospective likelihood can give biased estimate on the haplotype parameter due to overrepresentation of cases or the extremes of the traits compared with the general population. Therefore they proposed a retrospective model of haplotype frequencies conditional on the disease phenotype via logistic model. The model is applicable to a wide range of traits and provides unbiased estimates. A correspondence between retrospective and prospective risk estimates was also given.

A recent paper by Zou (2005) showed it can accommodate a variety of scenarios in genetic epidemiology, ranging from population designs, grouped or matched to family samples, individual or meta-analysis, for single SNP or haplotypes.

An example on polymorphism of the glycoprotein IIIa subunit of the fibrinogen receptor encoded by (ITGB3) and coronary heart disease (Weiss *et al.* 1996; Garcia-Ribes *et al.* 1998). The data are presented in the following table.

Two case-control studies on ITGB3 and CHD

	< 60			>=60		
	00	01	11	00	01	11
Weiss et al						
Case	21	19	2	22	4	3
Control	31	5	0	24	7	1
Garcia-Ribes et al						
Case	20	18	0	43	15	4
Control	41	3	0	46	9	1

```
data weiss_garcia;
input count@@;
cards;
21 19  2
31  5  0
22  4  3
24  7  1
20 18  0
41  3  0
43 15  4
46  9  1
data t;
  subject=1;
  do study="weiss","garciz-Ribs";
    do agelth60=1,0;
      do trait=1,0;
        do h=0,1,2;
          set weiss_garcia;
          select (h);
          when (0) do; allele=0; output; allele=0; output; end;
          when (1) do; allele=0; output; allele=1; output; end;
          when (2) do; allele=1; output; allele=1; output; end;
          end;
          subject=subject+1;
        end;
      end;
    end;
  end;
run;
proc print;
run;
proc genmod descending;
    class study subject;
    freq count;
    model allele=study agelth60 trait/dist=bin type3;
    repeated subject=subject / type=indep;
    estimate 'trait' trait 1/exp;
run;
```

The SAS codes yield the age-adjusted score statistic of 21.90 with $P = 2.87 \times 10-6$. The effect estimate is as follows,

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-square	P
trait	1.1104	0.2436	0.05	0.6329	1.5879	20.77	<.0001
Exp(trait)	3.0355	0.7395	0.05	1.8831	4.8933		

3. Other updates

The computer program hapstats is released from DY Lin's website (http://www.bios.unc.edu/~lin). He and others had several papers (Bioinformatics, Am J Hum Genet, Genet Epidemiol, J Am Stat Assoc, etc) on multiple testing, staged design, haplotype analysis in the context of a variety of designs such as case-control, nested case-control and case-cohort. This program perhaps can be seen as competitive to THESIAS (http://genecanvas.ecgene.net/downloads.php) and SimHAP (http://www.genepi.org.au/simhap).

A new package called GroupSeq is released with R 2.3.0 for group sequential designs. Meanwhile, Bioconductor is oriented to larger datasets based on a design using SQLite coupled with bounded memory algorithms.

4. Additional notes

Common to all aspects in mind here is the concept of integrated analysis of genetic data, or genetic data analysis in integrated systems. The steady increase in software development strengthens the idea of "publishing software, not just papers about software" and "reproducible research". There have been a lot of discussions in the literature and newly proposed systems, e.g. GENOMIZER (Franke et al. 2006) at Kiel of Germany, GMED developed at Boston University of USA, and similar system envisaged by Lon Cardon from WTCHG at Oxford, HelixTree (http://www.goldenhelix.com). However, the bioinformatics side has far-reaching impact, e.g. recent circulation from Affymetrics about T1D at CIMR.

The retrospective methods remain to be an interesting idea to pursue due to its advantage in power and versatility. It is worthwhile to apply them in our own data analysis.

References

Epstein MP, Satten GA. *Am J Hum Genet* **73**:1316–1329, 2003
Franke A, et al.. *Hum Mut.* **27**:583-588,2006
Garcia-Ribes M, et al. *Thrombosis and Haemostasis* **79**: 1126–1129, 1998
Lin DY, Zeng D. *JASA* **101**: 89-104, 2006
Prentice RL. *Biometrics* **32**:599-606, 1976
Satten GA, Epstein MP. *Genetic Epidemiol* **27**:192–201, 2004
Schaid DJ, et al. *Am J Hum Genet* **70**:425–434, 2002
Tan Q, et al. *Genet Res.* **86**: 223–231, 2005
Weiss EJ, et al. *NEJM* **334**: 1090-1094, 2001
Zou GY. *Ann Hum Genet* **70**:262–276, 2006
Zhao JH, Tan Q. *Curr Bioinformatics* **1**, 2006